Predicting the Position of Attributive Adjectives in the French NP

Gwendoline Fox, Juliette Thuilier

Predicting the Position of Attributive Adjectives in the French NP

Gwendoline Fox[1] and Juliette Thuilier[2]

[1] University of Paris 3 - Sorbonne Nouvelle (ILPGA) and EA 1483
[2] University of Paris 7 - Denis Diderot (UFRL) and ALPAGE (INRIA)

1 Introduction

French displays the possibility of both pre-nominal and post-nominal ordering of adjectives within the noun phrase (NP)[3].

(1) un magnifique tableau / un tableau magnifique
a magnificent painting / a painting magnificient
"a magnificient painting"

(2) un beau tableau / ??un tableau beau / un tableau très beau
a nice painting / a painting nice / a painting very nice
"a nice painting"

The above examples show that the positionning of attributive adjectives is a complex phenomenon: in (1) the adjective *magnifique* may be in both positions while *beau* strongly prefers anteposition, unless it is modified by an adverb, as seen in (2).

The question of adjective alternation has led to many studies in French linguistics ([1,2,3,4,5,6] among others). The constraints playing a role in this phenomenon are said to be phonological, morphological, syntactic, semantic, discursive and also pragmatic. Only one of the proposed constraints is categorical in the sense that it imposes a specific position to an attributive adjective: the presence of a post-adjectival complement (3) or modifier (4) only allows postposition of the adjective.

[3] The position of the adjective can imply semantic change, for specific adjectives (i) or specific noun-adjective combinations (ii):

(i) un coffre ancien vs. un ancien coffre
a chest old a old chest
" an old chest vs. a former chest"

(ii) un gros fumeur / un fumeur gros / un gros chanteur
a big smoker a smoker big a big singer
"an heavy smoker / a fat smoker / a fat singer"

We do not take into consideration these kind of semantic changes in this article. Our work focuses on the form of the adjective.

(3) un homme fier de son fils / *un fier de son fils homme
a man proud of his son / a proud of his son man
"a man proud of his son"

(4) un entretien long de deux heures / *un long de deux heures entretien.
a interview long of two hours / a long of two hours interview
" a two hours long interview "

The other constraints participating in the alternation between anteposition and postposition are not categorical. For instance, as noted in the corpus studies of [1,2], length and frequency are preferential constraints: short adjectives, as well as the most frequent ones, tend to be anteposed.

In order to account for preferential constraints, we present along the same lines as [7], a quantitative study of the position of attributive adjectives, based on two corpora: the French Tree Bank (henceforth FTB) and the Est-Républicain corpus (henceforth ER). The aim of this article is to propose a prediction model based on interpretable constraints and to compare their prediction capacities in order to better estimate their respective contribution in the choices guiding the placement of adjectives.

2 Methodology

Building the datatable The first step of this work is to collect the data concerning adjectives and capture the constraints found in the literature. The study is based on the functionnally annotated subset of the FTB corpus [8][4], which contains 12351 sentences, 24098 word types and 385458 tokens. It is, for the moment, the only existing treebank for French. We extracted all the occurrences of attributive adjectives from this corpus[5], and filtered out numeral adjectives[6], adjectives appearing in dates[7], abbreviations[8] and incorrectly annotated occurrences. We also discarded the 438 adjectives occurring with a post-adjectival dependent since postposition is imposed (see (3) and (4)) by a categorical constraint that overrides any other preferential constraint. The remaining adjectives constitute the basis of the datatable, to wich we have added information on the position of each adjective with respect to the noun it modifies, and 11 other variables that we describe in section 3.

Three variables of our study (FREQ, COLLOCANT and COLLOCPOST) were extracted from the ER corpus for more reliable counts. The raw corpus is a 147,934,722 tokens corpus, available on the ATLIF website[9]. It was tagged and

[4] This subset corresponds to the part that was manually corrected.
[5] We identified attributive adjectives using the following pattern in the treebank : an adjective occuring with an nominal head within a NP is an attributive adjective.
[6] Cardinal numerals such as *trois* 'three', *vingt* 'twenty', *soixante* 'sixty'... are sometimes annotated as adjectives in the FTB.
[7] Examples of dates containing adjectives : "$[13]_{ADJ}$ $[mars]_N$", "$[lundi]_N$ $[31]_{ADJ}$".
[8] Nouns or adjectives are viewed as abbreviations if their last letter is a capital letter.
[9] *http://www.cnrtl.fr/corpus/estrepublicain/*

lemmatized with the *Morfette* system [9] adapted for French. This corpus was used to compute the frequency of every adjectival lemma as well as Adjective-Noun and Noun-Adjective collocations.

The datatable contains 14804 occurrences corresponding to 1920 adjectival lemmas. 4227 (28.6%) tokens appear in anteposition, and 10577 (71.4%) in postposition. Table 1 shows that the adjectival lemmas displaying position alternance represent only 9.5% of all lemmas, yet these few lemmas correspond to 5473 occurrences, i.e. 37.0% of the datatable, which means that they are highly frequent adjectives.

Note that among the alternating adjectives (occurring in both positions), the ratio between anteposed and postposed occurrences is the reverse from that of all adjectives: there are 3727 anteposed (68,1%) and 1746 postposed (31,9%) adjectives. Alternating adjectives thus show a preference for anteposition. The general pattern is therefore that postposed adjectives tend to be infrequent lemmas occurring only in postposition, whereas alternating adjectives tend to be frequent and to prefer anteposition.

	anteposed	*postposed*	*both positions*	*Overall*
number of lemmas	125	1613	182	1920
	6.5%	*84.0%*	*9.5%*	*100%*
tokens	500	8831	5473	14804
	3.4%	*59.7%*	*37.0%*	*100%*

Table 1. Distribution of adjectival lemmas and tokens according to position

Statistical inference and logistic regression We used logistic regression models [10] in order to best estimate the distribution of adjective positions using the variables from the datatable. Formally, a logistic regression is a function for which values can be interpreted as conditional probabilities. Its analytical form is as follows:

$$\pi_{\text{ante}} = \frac{e^{\beta \mathbf{X}}}{1 + e^{\beta \mathbf{X}}} \tag{1}$$

where, in our case, π_{ante} is the probability for the adjective to be anteposed and β corresponds to the abbreviation of the sequence of regression coefficients α, $\beta_0 ... \beta_n$, respectively associated with the predicting variables $X_0 ... X_n$. Given a scatter plot, the calculation of regression consists in the maximum likelihood estimation of α and β_i parameters for each variable in a *logit* space.

This type of modelling consists in the combining of several explicative variables (binary or continuous) to predict the behaviour of a single binary variable, here the position of the adjective. More precisely, we estimate the probability of anteposition as a function of 11 variables. Given one adjectival occurrence and the value of the 11 explanatory variables attributed to this occurrence, the model predicts postposition if the probabilty is lower than 0.5, and anteposition if the probability is higher or equal to 0.5.

In order to evaluate the relevance of the constraints, we compare prediction models based on different constraint clusters. We use a 10-fold cross-validation to compute the accuracy of each model (noted μ and its standard deviation σ). The accuracy represents the proportion of data that is correctly predicted.

The comparison of the different models takes as a reference the accuracy of the *baseline model*: $\mu = 71.4\%(\sigma = 0.019)$. This model does not contain any explanatory variables and systematically predicts postposition. Its accuracy thus corresponds to the proportion of postposed adjectives in the datatable.

3 Variables

The variables we use in our logistic regression models are derived from the constraints found in the literature on attributive adjectives in French. They are summarized in table 2. Each model is based on different sets of constraints according to specific properties. The first set (COORD and ADV) concerns the syntactic environment of the adjective, the second is based on length and frequency (ADJ-LENGTH, AP-LENGTH and FREQ), the third one on the lexical properties of the adjectival item (DERIVED, NATIO, COLOUR and INDEF). Finally, the fourth group examines collocationnal effects of the Noun - Adjective combination (COLLOCANT and COLLOCPOST).

Variables	Types	Description
COORD	*bool*	adjective in coordination or not
ADV	*bool*	adjective with pre-modifying adverb or not
ADJ-LENGTH	*real*	length of the adjective in syllables
AP-LENGTH	*real*	length of the AP in syllables
FREQ	*real*	adjective frequency in the ER corpus
DERIVED	*bool*	derived adjective or not
NATIO	*bool*	adjective of nationality or not
COLOUR	*bool*	adjective of colour or not
INDEF	*bool*	indefinite adjective or not
COLLOCANT	*real*	score for the adjective-noun bigram
COLLOCPOST	*real*	score for the noun adjective bigram

Table 2. Summary table of variables and their values (*bool* = boolean and *real* = real number)

*Coordination (*COORD*)* In a competence account of attributive position like in [5], the position of coordinated adjectives is not restricted, as can be seen in example (5) (from [5]).

(5) une belle et longue table / une table belle et longue
a beautiful and long table / a table beautiful and long
"a long and beautiful table"

However, 94.6% of coordinated adjectival occurrences (i.e. 758 occurrences) are postposed in our data. Usage-based data thus suggests that coordination is a factor that strongly favours postposition.

*Presence of a premodifying adverb (*ADV*)* The general constraint is the same as for coordination: the presence of a pre-adjectival modifier does not restrict the position of the modified adjective (example (6)).

(6) une très longue table / une table très longue
a very long table / a table very long
"a very long table"

[5] point out that adjectives can be postposed with any adverb whereas only a small set of adverbs allows anteposition. This is confirmed in our datatable: 11 types of adverb[10] are observed with anteposed adjectives, while 119 different types appear with adjectives in postposition. Furthermore, the adverbs found with adjectives in anteposition are not specific to this position, they also appear with postposed occurrences. From a general quantitative point of view, 74.9% of the premodified adjectival occurrences are in postposition.

Length Numerous works on word order use the notion of length: for attributive adjectives in French [1,2], for word [11,12] and constituent [13,14,15,7] alternation in other languages. The main idea is expressed by the principle *short comes first*, i.e. short elements tend to appear first. Here, we consider length in terms of number of syllables and we introduce two variables: length of the adjective (ADJ-LENGTH) and length of the adjectival phrase (AP) (AP-LENGTH)[11].

*Lemma frequency (*FREQ*)* In his corpus study, [2] observes that high frequency is correlated with anteposition. In this work, we built a dictionary of frequency for each adjectival lemma in the ER corpus. We consider that frequency in ER better estimates the probability of use of an adjective than frequency in FTB, given the importance of the data (almost 1.5 million words for ER vs. 385,000 words for FTB)[12].

*Derived adjectives (*DERIVED*)* Adjectives may be derived from other parts-of-speech: e.g. from verbs (past participles, present participle, by suffixation: *-ible* 'faillible' (*faillible)* / *-able* 'faisable' (*doable*) /*if* 'attractif' (*attractive*) or from

[10] The 11 adverbs are : 'encore' *again*, 'désormais' *from now on*, 'moins' *less*, 'peu' *not much*, 'plus' *more*, 'si' *so*, 'tout' *very*, 'très' *very*, 'trop' *too*, 'bien' *well*, 'aussi' *also*.

[11] We obtain the number of syllables using the speech synthesis software ELITE [16]. It counts the number of syllables for every token, taking into account the actual form of the adjective (feminine versus masculine, for instance) as well as the possible effects of sandhi phenomena, like the *liaison* phenomenon. The value associated to each adjectival type corresponds to the mean of all its tokens length.

[12] The FREQ value of an adjective is 0 if the adjective is in the datatable but not in the ER corpus. Frequency being considered as a mere estimator, the model will handle such data similarly to very low non-null values.

nouns ('métallique' (*made of metal*), 'scolaire' (*academic*), 'présidentiel' (*presidential*)). These adjectives are described as prefering postposition. We marked them with the variable DERIVED[13].

Lexico-semantic classes Most reference grammars state that objective adjectives (i.e. adjectives for which the semantic content is perceptible or can be infered from direct observation) are postposed. Objective adjectives are classified into sub-groups like form, colour, physical property, nationality, technical terms... In order to estimate the relevance of lexico-semantic classes for the placement of adjectives, we test the predictive capacity of two classes by means of two variables: NATIO for nationality[14] and COLOUR for colour[15].

We also added the class of indefinite adjectives. These adjectives are special in the fact that their syntacitc properties show a hybrid behaviour between determiners and adjectives. On the one hand, indefinite adjectives may introduce and actualise the noun, like determiners. On the other hand, they may co-occur with a determiner and can be placed in post-nominal position, even though they favour anteposition. These latter properties are specific to attributive adjectives. The adjectives we identified as indefinite in our datatable are: 'tel' (*such*), 'autre' (*other*), 'certain' (*some/sure*), 'quelques' (*few*), 'divers' (*various*), 'différent' (*different*), 'maint' (*numerous*), 'nul' (*null/lousy*), 'quelconque' (*any/ordinary*), 'même' (*same/itself*). They are marked by means of the variable INDEF.

Collocations It is well known that the nature of some Adjective-Noun combinations is strongly collocational. This implies that the position of attributive adjectives in French should also be influenced by collocational effects. The collocation score in our datatable corresponds to the frequency of the Adjective-Noun (COLLOCANT) and Noun-Adjective (COLLOCPOST) bigrams in the ER corpus. We use raw frequency relying on the idea that the frequency of stored elements directly affects the representation of these elements [19]. As a further support, the experiment conducted by [20] shows that judgements elicited from human subjects about adjective-noun pairs in English are highly correlated with the co-occurrence frequency[16]. This suggests that frequency is a good association measure of collocational bigrams.

---

[13] The adjectives derived from another part-of-speech (noun or verb) are collected using the software of derivational morphological analysis DERIF [17].

[14] Using the dictionnary PROLEXBASE [18].

[15] Using the dictionnary CHROMA: *http://pourpre.com/chroma/*.

[16] The authors pointed out that frequency has the best correlation score compared to other association scores: conditional probability of the noun given the adjective, log-likelihood ratio, selectional association measure.

$$\pi_{\text{ante}} = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}, \text{ where}$$

$$\begin{array}{rrll} X\beta = & +0.86 & & *** \\ & -0.53 & \textsc{Coord} = 1 & * \\ & -0.55 & \textsc{adj-length} & *** \\ & -0.41 & \textsc{ap-length} & *** \\ & +0.00003 & \textsc{Freq} & *** \\ & -0.47 & \textsc{derived} = 1 & *** \\ & +1.74 & \textsc{Indef} = 1 & *** \\ & -5.25 & \textsc{natio} = 1 & *** \\ & -15.05 & \textsc{colour} = 1 & \\ & +0.003 & \textsc{CollocAnt} & *** \\ & -0.003 & \textsc{CollocPost} & *** \end{array}$$

Fig. 1. Formula of prediction model, significant effects are coded *** p<0.001, ** p<0.01, * p<0.1

4 Prediction model of attributive adjective position

The prediction model is built with all the variables described in part 3 and maximized with a backward elimination procedure based on AIC criterion [21][17]. The ADV constraint's contribution to the model is not significant according to the procedure. It was thus eliminated. The model is presented in figure 1.

As we expected, the variables COORD, ADJ-LENGTH, AP-LENGTH, DERIVED, NATIO, COLOUR and COLLOCPOST tend to favour postposition, whereas FREQ, INDEF and COLLOCANT vote for anteposition.

Compared to the *baseline model* performances ($\mu = 71.4\%, \sigma = 0.019$), this model has a significantly better accuracy ($\mu = 88.6\%$, $\sigma = 0.01$). The prediction performances are presented in table 7.

In order to compare the effect of different constraint clusters, we built 4 prediction models based on different groups of variables: a *Syntactic model* containing COORD[18]; a *Lexical property model* with NATIO, COLOUR, INDEF and DERIVED; a *Frequency-Length model* containing the variables ADJ-LENGTH, AP-LENGTH and FREQ and a *Collocations model* containing COLLOCANT and COLLOCPOST.

*Syntactic model (*COORD*).* First, the comparison shows that the effect of the syntactic constraint COORD is insignificant when it is not combined with other constraints (*Syntactic model* accuracy: $\mu = 71.4, \sigma = 0.02$). Moreover, the elimination of the ADV constraint from the prediction model strengthens the idea that syntactic constraints have no important predictive power. This can be partly ex-

[17] Note that for this particular model forward selection procedure gives the same results.

[18] We do not integrate the ADV constraint in the *Syntactic model* because the elimination with the AIC procedure already shows its lack of predictive power within the global model.

		Predicted position		% Correct
		P	A	
observed	P	10208	369	96.5%
position	A	1323	2904	68.7%

Overall accuracy: $\mu = 88.6\%$ ($\sigma = 0.001$)

Table 3. Classification table for *prediction model*

plained by the fact that these two variables are relevant for a very small set of data: ADV and COORD represent respectively 5.2% and 5.4% of all the data.

		Predicted position		% Correct
		P	A	
observed	P	10574	3	99.9%
position	A	4227	0	0%

Overall accuracy: $\mu = 71.4$ ($\sigma = 0.02$)

Table 4. Classification table for *Syntactic model*

*Lexical properties model (*NATIO, COLOUR, INDEF *and* DERIVED*).* Second, lexical properties are relevant when they are not combined with the other constraints (*Lexical properties model* accuracy: $\mu = 74.7, \sigma = 0.02$). This observation encourages us to extend the number of semantic classes in order to improve our modelling.

		Predicted position		% Correct
		P	A	
observed	P	10506	71	99.3%
position	A	3681	546	12.9%

Overall accuracy: $\mu = 74.7$ ($\sigma = 0.02$)

Table 5. Classification table for *Lexical properties model*

*Frequency-Length model (*ADJ-LENGTH, AP-LENGTH *and* FREQ*).* Third, we note that the variables of length and frequency have the most important predictive power (*Frequency-Length model* accuracy: $\mu = 85.8, \sigma = 0.009$). It is interesting that both variables are said to be led by processing ease. Length is positively correlated with complexity, be it from an articulatory point of view [22,23] or a syntactic one [14,24]. Based on the idea that the general processing cost is reduced when less complex elements precede more complex ones, it is expected that short adjectives should favour anteposition. Likewise, as [19]

argues, each use of a word is stored in the memory of locutors and added to the mental representation that they have of it. Each occurrence reinforces the mental representation and makes it more accessible for the locutor to process. In other words, a highly frequent item is highly accessible, and thus easy to process. Consequently, highly frequent adjectives are also expected to be anteposed for the general processing ease of the NP. The importance of both constraints in the prediction may be viewed as a support for these claims[19].

		Predicted position		% Correct
		P	A	
observed	P	10122	455	95.7%
position	A	1654	2573	60.9%

Overall accuracy: $\mu = 85.8$ ($\sigma = 0.009$)

Table 6. Classification table for *Frequency-Length model*

*Collocations model (*CollocAnt *and* CollocPost*).* Fourth, the *Collocations model* shows that the frequency of bigrams also represents a good predictor (*Collocations model* accuracy: $\mu = 79.7, \sigma = 0.013$). This observation is linked to the above-mentioned idea that frequency of use has an effect on mental representations. Indeed, as in [19] and in works on construction grammar [25,26], we may hypothesize that units larger than words are stored.

		Predicted position		% Correct
		P	A	
observed	P	10495	82	99.2%
position	A	2930	1297	30.7%

Overall accuracy: $\mu = 79.7$ ($\sigma = 0.013$)

Table 7. Classification table for *Collocations model*

To conclude this section, note that a large proportion of constraints playing a significant role in the studied phenomenon relates to the adjectival item (adj-length, Freq, natio, colour, Indef and derived). The choice of position thus seems largely determined by the use of a specific item.

[19] Note that frequencies are biased by the journalistic nature of corpora: adjectives of nationality are frequent despite the fact that they are postposed in most cases. Nevertheless, the variable natio of the global prediction model votes for postposition, which neutralizes the frequency effect.

5 Conclusion

We examined the question of position alternation of attributive adjectives in French using quantitative methods applied to corpora. One can draw several conclusions from the logistic regression models that we proposed. First the satisfactory results of our general model show that a good part of the modeling can be done on the basis of the form without considering the semantics due to position. Moreover, this work points out that lexical properties, including semantic classes, are relatively good predictors. These conclusions suggest that our knowledge on the very nature of adjectival items plays an important role in their positionning. Nevertheless, the prediction performances may be improved by taking more semantics into account: adding information for other semantic classes should naturally enhance the model. The question however remains on how to capture and formalise semantic relations in a quantitative study. Finally, our work outlines the importance of length and frequency-based constraints. This confirms the role of the nature of adjectives, but it also shows that usage strongly participates to the building of linguistic knowledge, and hence to the positionning of adjectives.

References

1. Forsgren, M.: La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique. Almqvist & Wilksell, Stockholm (1978)
2. Wilmet, M.: La place de l'épithète qualificative en français contemporain : étude grammaticale et stylistique. Revue de linguistique romane **45** (1981) 17–73
3. Delbecque, N.: Word order as a reflexion of alternate conceptual construals in french and spanish. similarities and divergences in adjective position. Cognitive Linguistics **1** (1990) 349–416
4. Nølke, H.: Où placer l'adjectif épithète? focalisation et modularité. Langue française **111** (1996) 38–57
5. Abeillé, A., Godard, D.: La position de l'adjectif èpithète en français : le poids des mots. Recherches linguistiques de Vincennes **28** (1999) 9–32
6. Noailly, M.: L'adjectif en français. Ophrys (1999)
7. Bresnan, J., Cueni, A., Nikitina, T., Baayen., H.: Predicting the dative alternation. In Boume, G., Kraemer, I., Zwarts, J., eds.: Cognitive Foundations of Interpretation. Royal Netherlands Academy of Science, Amsterdam (2007)
8. Abeillé, A., Barrier, N.: Enriching a french treebank. In: Proceedings of Language Ressources and Evaluation Conference (LREC), Lisbon (2004)
9. Grzegorz Chrupala, G.D., van Genabith, J.: Learning morphology with morfette. In Calzolari, N., *et al*, eds.: Proceedings of LREC'08, Morocco, ELRA (2008)
10. Agresti, A.: An introduction to categorical data analysis. Wiley interscience (2007)
11. Cooper, W.E., Ross, J.R.: World order. In: Papers from the Parasession on Functionalism. Chicago Linguistic Society (1975) 63–111
12. Benor, S.B., Levy, R.: The chicken or the egg? a probabilistic analysis of english binomials. Language **82(2)** (2006) 28–55
13. Hawkins, J.: The relative order of prepositional phrases in english: Going beyond manner-place-time. Language Variation and Change **11** (2000) 231–266
14. Wasow, T.: Postverbal behavior. CSLI publications (2002)

15. Rosenbach, A.: Animacy versus weight as determinants of grammatical variation in english. Language **81(3)** (2005) 613–644
16. Beaufort, R., Ruelle, A.: elite : système de synthèse de la parole à orientation linguistique. In: Actes des XXVI journées d'études sur la parole, Dinard (2006)
17. Namer, F.: Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In: Traitement Automatique de la Langue Naturelle (TALN). (2002)
18. Tran, M., Maurel, D.: Prolexbase : Un dictionnaire relationnel multilingue de noms propres. Traitement automatique des langues **47**(3) (2006) 115–139
19. Bybee, J.: The emergent lexicon. In Gruber, M.C., Higgins, D., Olson, K.S., Wysocki, T., eds.: CLS 34: The panels, University of Chicago, Chicago Linguistic Society (1998)
20. Lapata, M., McDonald, S., Keller, F.: Determinants of adjective-noun plausibility. In: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95). (1999) 30–6
21. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6) (1974) 716–723
22. Cutler, A., Cooper, W.E.: Phoneme-monitoring in the context of different phonetic sequences. Journal of Verbal Learning and Verbal Behavior **18(4)** (1978) 497–508
23. Pinker, S., Birdsong, D.: Speakers' sensitivity to rules of frozen word order. Journal of Verbal Learning and Verbal Behavior **18(4)** (1979) 497–508
24. Hawkins, J.: Efficiency and Complexity in Grammars. Oxford University Press (2004)
25. Croft, W.: Radical Construction Grammar. Oxford University Press (2001)
26. Goldberg, A.: Constructions at Work: the nature of generalization in language. Oxford University Press (2006)